FR-03-65

HumRRO

Human Resources Research Organization

# Third-Party Checking of 2003 Scaling and Equating for the Kentucky Core Content Test

Arthur A. Thacker
R. Gene Hoffman
Emily Dickinson-Bacci

Human Resources Research Organization (HumRRO)
950 Breckenridge Lane, Suite 170
Louisville, KY 40207
Phone (502) 721-9045
FAX (502) 721-9983

*Prepared for:*

Kentucky Department of Education
Capital Plaza Tower, 18th Floor
500 Mero Street
Frankfort, KY 40501

September 2003

# Third-Party Checking of 2003 Scaling and Equating for the Kentucky Core Content Test

## Table of Contents

## Summary

CTB and HumRRO independently calculated the scaled and equated raw-score-to-scale-score tables for the 2003 Kentucky Core Content Test. From those tables, cut points were identified that can be used to (1) assign student performance classifications (Novice, Apprentice, Proficient, or Distinguished (NAPD)) and (2) convert to school accountability indexes. Decisions regarding the handling of problem test items were discussed between CTB and HumRRO and in all cases both groups reached consensus. Results calculated by HumRRO were nearly identical to those calculated by CTB. Given that our scaling and equating results were nearly identical (small differences due to rounding, etc., that would not affect any students NAPD classifications) with those of CTB, we are assured that CTB did not commit processing errors.

# Third-Party Checking of 2003 Scaling and Equating for the Kentucky Core Content Test

## Introduction

Every year, the Kentucky Core Content Test (KCCT) [1] is scaled and equated by Item Response Theory (IRT) using a calibration sample of students in designated grades (4, 5, 7, 8, 10, and 11). Scaling involves the estimation of item parameters for the current year's test. These item parameters are linearly transformed to a 325-800 point scale and equated with previous years' scales. The results of scaling and equating are then used to construct raw-score-to-scale-score tables for every KCCT test form. Cut points are also identified so that students' raw scores can be translated to performance categories: Novice, Apprentice, Proficient, and Distinguished (NAPD).

Scaling and equating are done for the following grade/subject combinations:

Grade 4 - Reading, Science
Grade 5 - Math, Social Studies, Arts & Humanities, Practical Living/Vocational Studies
Grade 7 - Reading, Science
Grade 8 - Math, Social Studies, Arts & Humanities, Practical Living/Vocational Studies
Grade 10 - Reading, Practical Living/Vocational Studies
Grade 11 - Math, Science, Social Studies, Arts & Humanities

As a quality control step, personnel at CTB and the Human Resources Research Organization (HumRRO) conduct scaling and equating analyses simultaneously and independently. Researchers at both companies compare results at several steps throughout the process. If a result between CTB and HumRRO is not identical, then procedures are reviewed until the issue is resolved and both staffs get the same outcome. This way, the complex sampling, item parameter estimation analyses, Stocking-Lord equating, raw-score-to-scale-score transformations, and cut point identifications are checked and verified by two, autonomous agencies.

The procedures used by HumRRO are outlined in detail below.

[1] The test in use before 1998 was the Kentucky Instructional Results Information System (KIRIS) test.

## Sample Identification and File Construction

The first step in performing the required analyses was to identify a calibration sample for each grade/subject and construct files formatted for use with CTB's IRT programs. The procedures for accomplishing this task changed radically from 2002 to 2003. Data Recognition Corporation (DRC) scored student test forms and provided student data files, as a subcontractor to CTB, for the past several years. Scoring was conducted by CTB, rather than through a subcontractor, for the first time in 2003. Data formats were altered a great deal due to this change. HumRRO was required to alter data processing protocols and SAS programs in order to create data sets usable in CTB's scaling and equating programs (Pardux and Flux).

The first anomaly in the data was discovered during this procedure. Previously, data were posted by grade/subject as completed. In 2003, the entire data set was posted as two files, one for hand-scored and one for machine-scored data. Initial results for grades 7 and 11 were labeled as "suspect" so those grades could only be completed after a reposting of these files. Early data included in these files for other grades were also corrected, requiring HumRRO to re-start the process.

Kentucky selects most of its student population for use in the calibration sample for scaling and equating. However, some students are purposefully exempted (a student who leaves the test form completely blank, for example). CTB has devised a set of rules for including students in the calibration file based on KDE's recommendations and the CTB file structure. HumRRO independently wrote a SAS program to apply those rules. CTB and HumRRO compared results at several stages during this procedure and most differences in the two sets of files were resolved. However, in several cases, CTB's and HumRRO's calibration samples were consistently different by a few students. In most cases the files were only different by one student, but in seventh grade, HumRRO's files included 17 students CTB's files did not. Early attempts to discover the difference in the manner in which the exclusion rules were applied failed. Due to the short time allotted for this procedure and the relatively large number of students included in Kentucky's calibration sample (> 40,000 for all grade/subjects) CTB and HumRRO decided to eliminate the anomalous students from the calibration files. So, by mutual decision, whichever calibration file was smaller became the "official" calibration sample. Students in either CTB's or HumRRO's files that were not common to both files were deleted from the calibration sample. HumRRO and CTB verified that the samples were identical prior to beginning IRT processing.

## Scaling and Equating Procedures

Item response data for all of the 2003 test forms were scaled using CTB's PARDUX program. Based on IRT, PARDUX uses a three-parameter logistic model for multiple-choice items and a two-parameter model for open-response items to estimate item parameters. Item parameters from both these models are eventually transformed to a single scale.

The equating process involves the application of the Stocking-Lord procedure to two different sets of anchor[2] item parameters: anchor item parameters from 2001 and anchor item

[2] Anchor items were designated on one form for each grade/subject on the 2001 KCCTs. The same anchor form was readministered in 2003 with all items intact and occurring in the same sequence as in 2001.

parameters from 2003. These two sets of parameters are on different metrics. The 2003 parameters are on a theta metric (-1 to +1 scale) and the 2001 item parameters are on the "Kentucky metric" (325 to 800 scale). Stocking-Lord produces transformation constants (M1 and M2) that are used to linearly transform the 2003 metric onto the 2001 metric. This transforms all the 2003 item parameters onto the 325 – 800 scale, which traces back to the original 1992 scale.

HumRRO's anchor file for fourth-grade science did not initially match CTB's file. In 2001, an item was dropped from the fourth grade science test, causing multiple versions of those parameters to be created. It is possible that CTB had initially misidentified the operational parameters from 2001. When the mismatch was discovered, CTB provided a second anchor file that did match HumRRO's parameters. However, when CTB's anchor file failed to function, it was discovered that some of CTB's anchor items had not been correctly numbered to match items from 2001 to 2003. Once the item numbers were correctly applied, HumRRO and CTB's anchor files matched.

One other issue should be mentioned with regard to the production of anchor files. CTB uses the Flux program and several "hand-steps" in order to create anchor files. HumRRO uses a SAS program written specifically to produce the anchor files. CTB's and HumRRO's anchor files contained slight differences in the last decimal place for several parameters. Investigation of the differences revealed that Flux truncates at the last decimal while HumRRO's SAS program rounds. HumRRO discovered this anomaly in 2002 and created anchor files with Flux and with the SAS program to investigate whether there were meaningful differences caused by these slight inconsistencies. There were none in 2002. In 2003, HumRRO used only the anchors created using the SAS program. Again, the slight differences in M1, M2, and in student scoring tables would have caused no differences in student or school classifications.

The final step in the process is to use CTB's FLUX program to create raw-score-to-scale-score conversion tables and identify the cut points for the performance categories. The slight variations in M1 and M2 did cause some small differences in CTB's and HumRRO's scoring tables, but never by more than one scale score point and in no instance did the differences affect student performance classifications.

## Scope of Third-Party Checking

In addition to doing a parallel analysis with CTB this year, HumRRO also conducted in-house, parallel analysis to accomplish scaling and linking for the 2003 data. The Processing Steps listed below, while adequate, are being improved each year to ensure greater accuracy, standardization, and efficiency. This year, because of the changes in the student data files HumRRO received, a large portion of HumRRO's efforts were dedicated to reading the new files and creating calibration files correctly formatted for use in IRT processing programs.

## Processing Steps

HumRRO took the following steps for each grade/subject tested:

1. Created anchor files (PARDUX *.anc) of multiple-choice test items that appeared on the anchor form. These anchor items were used to equate the 2003 test to the 2001

scale. The 2003 anchor files were created using 2001 parameter files for the matching forms.

2. Created working files (PARDUX *.RWO) from the calibration sample for the 2003 Kentucky Core Content Test. These files include both open-response and multiple-choice data.

3. Prepared control files (PARDUX *.ctl) which contain the constraints used for item parameter estimation, student proficiency estimation, maximum number of items, etc. The SAS program used to create *.rwo files included a routine to print out a control file.

4. Estimated parameters for Kentucky Core Content Test items using PARDUX.

5. Performed Stocking-Lord transformation using PARDUX. The results of this transformation include a slope and intercept constant for equating the 2003 Kentucky Core Content Test back to 2001.

6. Confirmed that the equating constants (M1 and M2) from Step 5 match those derived by CTB.

7. Created parameter files (FLUX *.par) for each test form for use in preparation of raw-score-to-scale-score tables. A special SAS program was written for this purpose.

8. Created files (FLUX *.hlk) containing the scale limits (325 and 800) and constants from the Stocking-Lord transformation. A special SAS program was written for this purpose.

9. Created raw-score-to-scale-score transformation tables for each form using FLUX.

10. Confirmed that the raw-score-to-scale-score transformation tables from Step 9 match those derived by CTB and verified cut points used to separate student performance into Novice (Non-performing, Middle, High)/Apprentice (Low, Middle, High)/Proficient/Distinguished categories.

## Results

After performing periodic checks with CTB as individual tests were scaled and equated, HumRRO and CTB reached near-exact agreement on the equating constants for all grade/subjects. Table 1 summarizes the results of this study. Grade and subject are identified for each test in the first two columns, respectively. The stage at which convergence occurred (if at all) is recorded in the third column. The fourth column identifies problem items and references the solutions that were reached by CTB and verified by HumRRO. The next four columns contain the M1 and M2 (slope and intercept) constants obtained from the Stocking-Lord transformation. CTB computed the first set of constants and HumRRO the second. The ninth column contains the difference between CTB's and HumRRO's M1 constants (i.e., $M1_{CTB}$-$M1_{HumRRO}$). The tenth column records the same information for M2 constants (i.e., $M2_{CTB}$-$M2_{HumRRO}$).

The last two columns in Table 1 list whether there was exact agreement between CTB and HumRRO on (1) the raw-score-to-scale-score tables and (2) the cut points. Cut points from these tables are used to assign students to performance categories that, in turn, are used in the computation of each school's accountability index. CTB and HumRRO were in near-exact agreement for all raw-score-to-scale-score tables for every grade/subject.

Explanations of convergence issues and individual item issues are footnoted in Table 1. The footnotes explain the specific problems and their solutions. It should be noted that all problem items were dealt with during the parameter estimation phase of the scaling and equating process. No item for which parameters were estimated was eliminated from the Stocking-Lord procedure. The same column indicates whether or not convergence was reached during parameter estimation. If convergence was not reached after 50 iterations by the PARDUX program, the solution at stage 50 was accepted by mutual agreement.

Table 1 includes three rows for grade five social studies. The first row indicates that item 126 was problematic. In fact, only two students scored the maximum four points for that item. Pardux uses a minimum cell size of three students in order to estimate parameters. The first row shows a solution created by recoding the item for a maximum score of three points rather than four. The second row shows a solution, suggested by CTB, which alters the minimum cell size to one and estimates parameters using only those two students. Neither of these solutions was used operationally and data for them are included in Table 1 in italics. After much discussion with CTB and Wested (a CTB subcontractor that helps write and create scoring guides for the test items) KDE decided to delete item 126 completely. The final line for grade five social studies contains data for this solution, which was used operationally for scoring.

**Table 1. KCCT 2003 Results**

| | | | | CTB | | HUMRRO | | CTB-HUMRRO Differences | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Grade* | *Subject* | *Convergence* | *Problems* | *M1* | *M2* | *M1* | *M2* | *M1* | *M2* | *Score Tables Agreement* | *NAPD Exact Agreement* |
| 4 | RD | stage 14 | none | 30.27457 | 550.56879 | 30.27497 | 550.56903 | -0.00040 | -0.00024 | Yes | Yes |
| | SC | stage 17 | none | 24.42577 | 551.52582 | 24.42596 | 551.52582 | -0.00019 | 0.00000 | Yes | Yes |
| 5 | A&H | stage 13 | none | 56.48065 | 524.85980 | 56.48091 | 524.85968 | -0.00026 | 0.00012 | No[3] | Yes |
| | MA | stage 17 | none | 34.99376 | 563.47272 | 34.99395 | 563.47278 | -0.00019 | -0.00006 | Yes | Yes |
| | PL | stage 15 | none | 44.58876 | 509.65042 | 44.58860 | 509.65030 | 0.00016 | 0.00012 | Yes | Yes |
| | *SS* | *stage 14* | *item 126* | *30.09034* | *543.82190* | *30.09048* | *543.82190* | *-0.00014* | *0.00000* | | |
| | *SS* | *stage 14* | *cell minimum* | *30.09060* | *543.82318* | *30.09063* | *543.82330* | *-0.00003* | *-0.00012* | | |
| | SS | stage 12 | Item 126 dropped | 30.13410 | 543.90948 | 30.13410 | 543.90948 | 0.00000 | 0.00000 | Yes | Yes |
| 7 | RD | none | convergence, item 135 M-step | 28.41569 | 516.28455 | 28.41588 | 516.28442 | -0.00019 | 0.00013 | No[4] | Yes |
| | SC | Stage 14 | none | 26.73661 | 505.36353 | 26.73663 | 505.36353 | -0.00002 | 0.00000 | Yes | Yes |
| 8 | A&H | stage 20 | item 70 & 79 M-step | 61.67681 | 520.40826 | 61.67686 | 520.40839 | -0.00005 | -0.00013 | Yes | Yes |
| | MA | stage 23 | none | 32.15652 | 537.85004 | 32.15650 | 537.85004 | 0.00002 | 0.00000 | Yes | Yes |
| | PL | stage 14 | none | 44.50098 | 506.65723 | 44.50481 | 506.65994 | -0.00383 | -0.00271 | No[5] | Yes |
| | SS | stage 16 | items 75, 83, 90, 134 M-step | 39.03960 | 517.16016 | 39.03965 | 517.16016 | -0.00005 | 0.00000 | No[6] | Yes |
| 10 | PL | stage 14 | none | 45.56385 | 507.34958 | 45.56414 | 507.34961 | -0.00029 | -0.00003 | No[7] | Yes |
| | RD | stage 17 | item 140 M-step | 50.52353 | 510.46954 | 50.52352 | 510.46964 | 0.00001 | -0.00010 | Yes | Yes |
| 11 | A&H | stage 19 | none | 56.24224 | 527.16895 | 56.24218 | 527.16888 | 0.00006 | 0.00007 | Yes | Yes |
| | MA | none | convergence | 40.71188 | 539.1037 | 40.7118 | 539.1037 | 0.00008 | 0.00000 | Yes | Yes |

[3] Form 1A is different in one instance by one scale score point.

[4] Forms 2A and 2B are each different in four instances by one scale score point. Forms 5A and 5B are each different in two instances by on one scale score point. Forms 6A and 6B are each different in one instance by one scale score point.

[5] Five forms (1B, 2A, 3A, 4B, 6A) are different in one instance by one scale score point.

[6] Forms 3A and 3B are each different in one instance by one scale score point.

[7] Form 4A is different in one instance by one scale score point.

**Table 1. KCCT 2003 Results**

| | | | CTB | | HUMRRO | | CTB-HUMRRO Differences | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| SC | Stage 16 | none | 32.04132 | 545.68341 | 32.04129 | 545.68341 | 0.00003 | 0.00000 | Yes | Yes |
| SS | stage 15 | items 98 & 104 M-step | 50.82182 | 547.31061 | 50.82172 | 547.31061 | 0.00010 | 0.00000 | Yes | Yes |

HumRRO also verified the cut points on the raw-score-to-scale-score tables. Cut points were assigned by rule. HumRRO verified cut points between Novice and Apprentice, between Apprentice and Proficient, and between Proficient and Distinguished performance categories. HumRRO also verified cut points for Low, Medium, and High subcategories within the Novice and Apprentice categories.

## Documentation

To document the steps involved in scaling and linking the 2003 Kentucky Core Content Test, HumRRO saved all electronic files used in data preparation, including SAS programs, SAS logs, and SAS output lists and all files produced during PARDUX scaling and FLUX transformations. These files have been submitted to the Kentucky Department of Education (KDE). Appendices from the Hoffman and Thacker (1999) report contain hardcopy examples of important files that were submitted.

All electronic files submitted to KDE are named according to the following code (where S = subject, G = grade level).

A. PARDUX Control File (SSGG03.CTL). This file contains the number of items, the maximum number of stages for PARDUX, the convergence criterion, parameter estimation limits, maximum and minimum values for proficiency estimates (theta). It also contains information allowing the program to distinguish between open-response and multiple-choice items, the number of score levels for open-response data, and which items to include in parameter estimation.

B. PARDUX Data File (SSGG03.RWO). This file contains the student score data. It is coded such that a 1 indicates a correct answer for a multiple-choice question and actual score levels (0-4) are recorded for student responses to open-response questions. To facilitate communication, HumRRO adhered to CTB's item order in constructing these data files.

C. PARDUX Anchor File (SSGG03.ANC). This file contains common-scaling item parameters from the 2001 KCCT (the identical items appeared on the 2002 KCCT). Only multiple-choice items are used in *.ANC files.

D. SAS Programs configured as SSGGrwcd.sas. This program produces the anchor files (*.ANC), PARDUX control files (*.CTL), and student score files (*.RWO). The SAS log and list files generated by these programs are also included electronically.

E. SAS Programs configured as SSGGmakeparfiles.sas. For each grade-subject, this program sorts the parameter data by test form, a configuration required by the FLUX program.

F. PARDUX Parameter Estimation Summary (SSGG03_SUM.TXT). This file provides a summary of the parameter estimation procedure run in PARDUX. It includes the limit data from the control file and also contains the number of stages PARDUX ran in order to reach convergence. It also contains the item numbers of items that could not be estimated and documents any items whose estimation reaches the maximum alpha parameter. This file identifies any problem items that might require additional manipulation before continuing the process.

G. PARDUX Parameter Estimation Details (SSGG03_DET.TXT). This file lists a systematic iteration of data, by item, during each stage of parameter estimation.

H. PARDUX Parameter File (SSGG03.PAR). This file contains parameter estimates for all items designated in the *.CTL file. It is used for later data manipulation.

I. PARDUX Item Summaries Files, Status (SSGG03_STAT.TXT). This file lists all items for a given test and their status after parameter estimation. Items are coded as either "estimate OK," "OK—default C," "not estimated," or "other codes." It provides a different type of record for the parameter estimation.

J. PARDUX Item Summaries Files, Distribution (SSGG03_DIST.TXT). This file contains the distribution of students who scored at each level on the open-response items. It is useful for examining the way that scoring rubrics for these items operate and for ensuring that all open-response items have the correct number of functioning score levels.

K. PARDUX Item Summaries Files, Parameters (SSGG03_PAR.TXT). This file contains the item parameters in different format from the *.PAR files. Word processing and spreadsheet programs can easily read this file.

L. PARDUX Item Summaries Files, Standard Errors (SSGG03_SE.TXT). This file contains the standard errors of estimation for each item including the errors for the various score levels on the open response items.

M. PARDUX Item Summaries Files, FitQ1 (SSGG03_Q1.TXT). This file contains fit statistics for all items.

N. PARDUX Log File (SSGG03_LOG.TXT). As each manipulation of data is completed, PARDUX maintains a log of the procedures and filenames. This log is saved in text format.

O. Stocking-Lord Plots (SSGG03_SLPLOTS.doc). For each grade/subject combination, the Stocking-Lord data transformation calculates M1 and M2 values (slope and intercept) and outputs four graphs (one each for the a, b, and c parameters, and item p-values). The M1/M2 values, a log of the Stocking-Lord procedures, and the graphs are saved in this file.

P. FLUX control file (SSGG03.HLK). This file specifies the range of the scale scores as well as the M1 and M2 transformation constants from the Stocking-Lord transformation.

Q. FLUX Parameter Files by Form (SSGG031A.PAR, SSGG031B.PAR, etc.). Each parameter file computed using PARDUX was divided to represent items from each test form. Typically, 30 items were scored from each form. The exceptions are forms from Arts and Humanities and Practical Living/Vocational Studies, which each contain only 10 scored items.

R. Raw-Score-to-Scale-Score Tables (SSGG03RStoSSTables.doc). A raw-score-to-scale-score table was produced for each form. These tables were saved in text format using FLUX.

S. Miscellaneous files and programs may also be included in the documentation. These files were constructed either during investigation of results or for future purposes. Student data records (provided by CTB) from which all 2002 data were extracted are included as well.

## Conclusion

CTB and HumRRO independently calculated the scaled/equated raw-score-to-scale-score tables for the 2003 Kentucky Core Content Test. From these tables, both identified cut points that could be used for assigning student performance classifications and later converted to school accountability indexes. No significant differences were found between CTB's and HumRRO's parameter estimation, Stocking-Lord transformation constants, raw-score-to-scale-score tables, or application of cut points. The differences that were found were in rounding of anchor item parameters – these rounding differences were so small that they had negligible effect on M1/M2 values and no effect on final cut points.

Given that the HumRRO and CTB scaling and linking results were nearly identical, HumRRO is confident that CTB did not commit processing errors.

## References

Hoffman, R. G. & Thacker, A. A. (1999). *Third-party checking of 1999 scaling and linking for the Kentucky Core Content Test*. (HumRRO Report SP-WATSD-99-44). Alexandria, VA: Human Resources Research Organization.